*3L: Language, Linguistics, Literature® The Southeast Asian Journal of English Language Studies*
*Vol 28(2), June 2022 http://doi.org/10.17576/3L-2022-2802-15*

# Lexical Richness of Abstracts in Scientific Papers in Anglophone and Non-Anglophone Journals

RODRIGO TOVAR VIERA
*Department of National and Foreign Language Pedagogy,*
*Technical University of Cotopaxi, Latacunga, Ecuador*
*rodrigo.tovar@utc.edu.ec*

ABSTRACT

*Journal research article abstracts, considered as the most read section of the entire paper, have been the focus of multi-dimensional research studies. In the genre of abstracts, vocabulary richness is the basis for the construction of sentences, paragraphs, and complete texts; it contributes to non-native (and even native) English speakers of the language in the production and comprehension of written texts. To explore its importance, the present study examines the lexical richness in abstracts of scientific papers; it consists of three distinct measurement dimensions: lexical density, lexical variation, and lexical sophistication. The comparative-descriptive analysis is based on a corpus of abstracts in English published in Anglophone and non-Anglophone contexts. The written corpora were subjected to a software-driven text analysis using the complete lextutor vocab-profile available online at https://www.lextutor.ca/vp/eng/; the output texts from the software analyser were mined using SPSS statistics. Results show that although abstracts in both publication contexts use varied and extensive vocabulary throughout the two English sub-corpora, Anglophone texts, unlike non-Anglophone ones, produce more content and off-list words. This study announces valuable insights, particularly for inexperienced and novice writers, on using automatic online tools, such as vocab-profile, to gauge the type of vocabulary used in their written compositions.*

*Keywords: abstract; lexical richness; publication context; text analyser; written corpora*

## INTRODUCTION

In linguistics, a lexicon is the inventory of lexemes of a language; lexemes are units of lexical meaning that correspond to a set of words connected through combination and formation. The branch of linguistics that examines lexicon or lexis, specified as the vocabulary or total stock of words of a language, is lexicology (Lipka, 2002). It drives from the Greek *lexikós* —of words and *logia* —study (Arnold, 1986, p. 272). Since the main goal is the systematization of words revealing characteristic features, it deals not only with lexemes and their properties but also with word combination and phraseological entries. Barbara Strang (1968), in her book Modern English structure, remarks that, unlike grammar, "lexis is the domain of vast list of formal items about which a rather little generalization can be made" (p. 215). Although the lexicon of a language is the accumulation of words, it is not constituted of a formal list of isolated elements. The lexicon results from various types of combinations and formations between the list of those interdependent lexical items, as *internal* (morphological relations, for example, affixation, compounds) and *external* (paradigmatic and syntagmatic relations, for instance, opposition, substitution). Then, it is assumed that there are multiple relations between lexicon and grammar choices when constructing written discourses. As a sort of remark on the lexicon structure, we can argue that the vocabulary of English (and other languages interpreted by different linguistic levels) is far from being homogeneous.

Although the lexicon is recognized as the network of lexemes (words) stored in the brain, including content and function words (McArthur et al., 2018), the entire vocabulary knowledge that a learner has acquired (Colman, 2015) consists of lexical entries that contain semantic and syntactic information about each item (Aarts, 2014). Lexical knowledge is the "progressive levels of knowledge, starting with a superficial familiarity with the word and ending with the ability to use the word correctly in free production" (Laufer et al. 2004, p. 400). Lexicon, therefore, accounts for the abundance of rich vocabulary in text production (Šišková, 2012) both in the mother tongue (L1) and in the non-native one, the latter either in the second (L2) or foreign language (FL). Such abundance of vocabulary is shown by gauging the lexical elements of density, diversity, and sophistication, which are indicators of lexical richness (Read, 2000; Laufer & Nation, 1995). In this way, vocabulary, the set of words used in writing (Colman, 2015), is essential for linguistic knowledge. Thus, according to Halliday and Hasan (1976), vocabulary is the core of meaning construction in discourse comprehension. Under this perspective, vocabulary can be defined by analysing the lexical size - how many words a learner knows; lexical depth - how well the learner knows the words; and fluency - the learner's ability to retrieve the form or the meaning of a word used in each context.

The structure of the English vocabulary, therefore, deals with the variety of the language use in the surrounding contexts, such as in the media, social classes, and academia. The latter, academic vocabulary, with the exponential number of scientific publications around the world, has been the common core of researchers and practitioners from different disciplines (e.g., Djiwandono, 2016; Durrant 2014; Romero, 2020; Šišková, 2012; Schmitt, 2000; Tovar, 2017) to find out regularities or differences throughout the texts. In this way, English and non-English speaking academic journals under well-classified standards attract a large number of research article (RA) submissions for analysis and possible publication. Apart from RA formats and policies, authors must summarize the scope of the papers in well-structured research article abstracts (RAAs) written in English. With the emergence of scientific production nationally and internationally, RA abstracts have become the most read research literature for annotated bibliography and catalogue reviews. This communicative function has made RAAs to be the foci of academic research in investigating their lexical richness (e.g., Bahtiar et al., 2020; Hung et al., 2021; Lin & Lin, 2019; Pho, 2008; Sánchez, 2020; Read, 2000: Tankó, 2017; Van Bonn & Swales, 2007; Waluyo & Kakoko, 2021;) from different dimensions. Some of these research studies have reported vocabulary variation, including diversity, sophistication, and density of the total words used in a text. In addition to those findings, Read (2000) argues that extensive vocabulary knowledge allows word diversification and avoids the repetition of related words. Since lexicon is one of the linguistic features involved in producing intelligible texts, "exploration into this area serves to map the ability to express ideas" rich in vocabulary (Djiwandono, 2016, p. 210). Based on these current research viewpoints, this study aims to describe the vocabulary size and its coverage in abstracts of scientific texts to find out common grounds or differences across disciplines between the two publication contexts.

## LEXICAL RICHNESS AND FURTHER RESEARCH

Lexical richness, the matter of this study, refers to writers' ability to effectively use lexical items in particular texts (Lewis, 1993). In the current research, lexical richness measures the word types and word diversity in abstracts of scientific papers. Accordingly, lexical richness is determined on

the dimensions of a) lexical variation (LV: diversity of words used), b) lexical sophistication (LS: specialized and academic words used), and c) lexical density (LD: proportion of content words used). LV describes the occurrence of different words in a text; it covers the relationship Type-token radio (TTR) and bases the analysis on the radio of different words (type) compared to the total number of words (token). LS relates to the quantity of complex, advanced, and specialized terminology presented in a text. LD refers to the proportion of lexical items (content words) related to the number of function words in each discourse. Linguistics and corpus linguistics research have considered these dimensions as one of the criteria to gauge vocabulary knowledge in writing compositions by using data-driven text parsers (e.g., Cobb, 2006). In the context of English as a Foreign Language (EFL), writing is challenging for those who are not directly exposed to and familiar with the conventions of the target language.

Researchers in corpus linguistics have documented the importance of vocabulary size and vocabulary knowledge in written corpora from distinct approaches (e.g., Nation, 2000, Nation, 2001, Schmitt, 2000, Laufer & Nation, 2001). Such research were reported to be valuable and reliable sources for further works in the field. Reviews on lexical richness revealed that almost native and non-native texts differ somewhat in lexical diversity, word frequency, lexical density, familiarity, and expressive vocabulary (e.g., Lee, 2018; Douglas, 2019; Ha, 2019). Factors influencing such variation could be the authors' linguistic background, cultural differences, and language proficiency. Read (2000) states that lexical richness, namely LV, LD, and LS, varies according to writing skills and language proficiency. Assessing the lexical richness of English texts, Laufer and Nation (1995) found out that the use of the first and second thousand words and sophisticated words vary across composition patterns. Aside from punctuation, grammatical accuracy, coherence, and cohesion, proper use of vocabulary makes writing easy to read and follows the central idea and argumentations. Laufer and Nation argue that "a well-used rich vocabulary is likely to have a positive effect on the reader" (p. 307). The incidence of various words shows that the writers have somehow exposed themselves to a wide range of reading materials from different types of texts (Djiwandono, 2016). Thus, lexical richness indices, in large part, are "relevant indicators of English writing proficiency level" and determiners of lexical knowledge to produce intelligible texts (Ha, 2019, p. 4); so that writing ability highly correlates with lexical knowledge and predicts academic success (Douglas, 2012). Then, it makes sense that knowing different words avoids repeating words and producing well-written compositions.

Investigating the lexical richness in academic papers, Djiwandono (2016) found variation in word diversity and academic words between experienced and inexperienced writers. Lexical richness, therefore, attempts to gauge the type and level of vocabulary used in different texts and the context of publication. Romero (2020) observed that EFL learners exhibited difficulties in understanding academic texts because their vocabulary level is less than 80%, which according to Hu and Nation (2000), conditioned their comprehension and writing production of the texts. In this way, Tovar (2017) points out that vocabulary knowledge is essential since it contributes to the comprehension and production of the foreign language. Although journal RAAs, written in English, have been extensively investigated, evidence from the research literature on this type of text is not substantial, particularly in Ecuador. To date, investigations in Ecuador examining the lexical richness of the RAAs are scarce. The scarcity of research on journal RAAs published in English and non-English-speaking contexts and how English abstracts from different disciplines and fields effectively use vocabulary to construct the content justifies the research interest. The present study describes the lexical richness of English RAAs written in education and electronics disciplines and published in Anglophone and non-Anglophone journals (henceforth, ANAJ).

Despite the state-of-the-art discussed in previous paragraphs, this study, as stated above, is still considered necessary because it gives continuity to the growing body of research carried out in this disciplinary area that until now has been explored by few researchers when examining the publication context of the text. Thus, the following research question drives the study:

To what extent do abstracts published in Anglophone journals show higher lexical dimensions than those published in non-Anglophone ones?

## METHODOLOGY

### DATA COLLECTION

The analysis is based on a corpus of 120 RAAs written in education and electronics disciplines, published between 2010 and 2018 in the Sciences and Humanities fields. The rationale for choosing education and electronics disciplines is because, according to Biber and Gray (2016), "the language of science research writing is quite different from the language of humanities prose." It claims the assumption that academic texts in different knowledge fields present different types of composing patterns. Table 1 below illustrates the characteristics of the two datasets of article abstracts in detail. Although there are considerable differences, for instance, in the mean of abstracts length, these differences do not affect this comparative-descriptive analysis. It is because the lexical richness counts the words that commonly occur from one text to another. The two English sub-corpora were compiled for the examination. Each corpus contains 60 RAAs selected from journals that follow the criteria used in Tovar (2019). First, RA published in the periods 2010 – 2018 were collected (288 RAAs: 144 in Anglophone (U.S.A) and 144 in non-Anglophone (Ecuador) journals). Second, the texts reporting a peer-review and editorial scrutiny were then considered; this was carried out to ensure that the comparative-descriptive analysis is a representative sample of the target language of abstracts written in education and electronics in Anglophone and non-Anglophone journals. Third, each corpus in ANAJ contains "abstracts selected from journals that meet the criteria of a) representativeness —*appropriate sample testing group*, b) reputation —*indexation and double-blind peer-reviewed*, and c) accessibility —*in print or online database.*" (p.77)

TABLE 1. Summary of the dataset

| | RAA | Abstracts length | | # of sentences per abstract | | Sentence length | | Total words |
|---|---|---|---|---|---|---|---|---|
| | | Mean | *SD* | Mean | *SD* | Mean | *SD* | |
| U.S.A. | 60 | 178.07 | *38.79* | 7.52 | *2.42* | 21.28 | *11.22* | 9923 |
| Ecuador | 60 | 168.27 | *48.16* | 6.11 | *2.36* | 34.44 | *11.28* | 9286 |
| Total | 120 | | | | | | | 19209 |

The research explored the education and electronics RAAs because, according to Ecuador's (2008) Constitution, these disciplines are the basis for innovation, promotion, development, and dissemination of knowledge and culture internationally. These areas are also the

professions with ample job opportunities (INEC, 2018) and interest in the higher education system; they are linked to global and technological production to generate scientific and technological research. Academics must spread their research findings to gain knowledge in their career and academic writing skills, mainly when writing abstracts of scientific papers in English.

## CORPUS SELECTION

Journal RA abstracts representing the "national or international indexing, double-blind peer-reviewed, unstructured text, and single paragraph condensed summary" were part of the corpus-based design criteria (Tovar, 2019, p. 77). Additionally, these RAAs should be published between 2010 and 2018 in Anglophone and non-Anglophone speaking contexts. Then, the publication context is the foci of the abstracts' selection rather than the nativeness of academic writers. This is because many speakers from different countries speaking distinct languages may use the same names that are usually common in English-speaking countries (Sayfouri, 2010).

The rationale behind choosing U.S.A. journals, such as the American Journal of Education (AJE), Journal of Teacher Education (JTE), Journal of Electronic Materials (JEM), Journal of Electronic Packaging (JEP), was their accessibility and the qualified record of the indexing process in Scopus. Such process included the degree of English language proficiency, rhetorical and writing style, knowledge of the disciplinary field, and the overall level of comprehensibility. Thus, the content quality and quantitative measures of the articles published in these journals are expected to be based on standard scientific English accepted by research and discourse communities. Additionally, these journals gained a respectable level of publication ranking and citation as domestic and international journals; they are indexed in some of the international databases to which non-anglophone journals belong. Such characteristics are positioning them as journals to guarantee controlled comparison analysis. It was, therefore, logical to expect that the hierarchy of these journals is the same as the content quality deployed in the sample RA abstracts. The data set of American journals includes 60 RAAs, 30 in each discipline. These journals are indexed in ERIC, EBSCOhost, Elsevier, American statistical association, SciSearch, SCOPUS, Applied Science and Technology, and Google Scholar.

Correspondingly, the Ecuadorian corpus consists of 60 abstracts written in English and published in Ecuador in the following journals: Alteridad (journal of education), UTCiencia (science and humanities), Revista Tecnológica ESPOL (science & humanities), Sophia (philosophy & education), ACI Avance (science & engineering), Enfoque (scientific engineering journal), Ingenius (science and technology), Maskay (electric & electronics). It encompasses 30 abstracts in each discipline. Ecuadorian journals report indexing in latindex (regional cooperative online information system for scholarly journals from Latin America, the Caribbean, Spain, and Portugal) and an international database, like Elsevier, DOAJ, Dialnet, REDIB, DRJI, SIS, MIAR, SciELO, and EBSCOhost. The higher education system hosts Ecuadorian journals. Unlike American journals, Ecuadorian ones are mixed, which devote special sections and space for the disciplines mentioned earlier. Thus, sample articles from Ecuadorian journals were expected to employ the informational sections of their abstracts precisely and skilfully per se compared to those from American journals.

PROCESS OF NARROWING THE CORPUS COLLECTION

After selecting journals from each discipline, a corpus of 288 abstracts was listed: 144 texts from Anglophone journals and 144 from non-Anglophone ones. Both Anglophone and non-Anglophone journals included 72 abstracts per discipline, like education and electronics. Using the *stattrek stratified random sampling* selection program (Harvey, 2000), which is available online at http://stattrek.com/Tables/Random.aspx, out of 288 texts, we obtained the final 120 RAAs: 60 abstracts in Anglophone journals and 60 ones in non-Anglophone journals. To give the *strattrek* program functionality, abstracts, in a spreadsheet, were first numbered from 1 to 144 with their respective discipline and publication context. Then, the *stattrek stratified random sampling* function was activated. Random sampling was carried out four times to select thirty RAAs from each two disciplines in ANAJ journals.

DATA ANALYSIS AND INSTRUMENTS

The two English sub-corpora published in ANAJ underwent software-based text analysis. With the top-down and bottom-up approach, sentences were the unit analysis of the study. The top-down approach focuses on the information content while the bottom-up approach looks for linguistic signals. The complete lextutor vocab-profile (Cobb, 2006), available online at https://www.lextutor.ca/vp/eng/ examines the lexical richness, namely lexical density (LD), lexical variation (LV), and lexical sophistication (LS) of journal RAAs. The two English sub-corpora were pasted to the submit window (Cobb), which performs a lexical text analysis of the written compositions. The output text displays the percentage of content words compared to the function words (LD), the diversity of words used in the text (LV), and the percentage of low-frequency words and off-list words (LS), which according to Laufer and Nation (1995), constitute the lexical richness of any text.

The output texts of the software analyser were mined using SPSS Statistics to determine the lexical richness of the texts. Accordingly, after obtaining the lexical dimensions for each English-sub corpus, a set of independent sample t-tests were run to compare and contrast RAAs published in ANAJ. Additionally, one-way ANOVAs were run to compare the two non-Anglophone written texts against the Anglophone groups. The Vocab-profile computer program was chosen because of its accessibility, high reliability, and extensive use as an automatic text analyser in a large set of linguistic research.

## RESEARCH FINDINGS

Data in Table 2 and figure 1 present the descriptive statistics of the average first and second 1000 most frequent words, non-repeated words, academic and uncommon words, and proportion of content words used in the abstracts published in ANAJ. The K2 word level in non-Anglophone journals (73.178) is slightly higher than that of the Anglophone journals (71.300). In this analysis, both English sub-corpora produced less than 80% of the K2 word list appearing in English texts (Nation, 2001). The type-token radio or lexical variation in Anglophone groups (0.624) is slightly higher than non-Anglophone ones (0.601). Anglophone texts, unlike non-Anglophone ones, produce more content and off-list words. This lexical difference turns low-frequency words into high-frequency words in a particular context where common and considered easy words might not

be familiar in the academic context (Gardner, 2013). Words that are not part of any family in English may be among the high-frequency words used in specialized texts and represent the use of advanced vocabulary and specialized terminology.

TABLE 2. Mean value indices of the lexical richness

| Variable | 1000 words (K1) | 2000-words (K2) | Lexical Variation (TTR) | Academic Words (LS) | Off-list word | Content word (LD) |
|---|---|---|---|---|---|---|
| | Mean | Mean | Mean | Mean | Mean | Mean |
| NAJ | 68.523 | 73.180 | 0.601 | 87.422 | 11.990 | 0.597 |
| AJ | 64.942 | 71.300 | 0.624 | 83.902 | 16.090 | 0.637 |

Note: *NAJ= non-Anglophone journals; AJ=Anglophone journals*

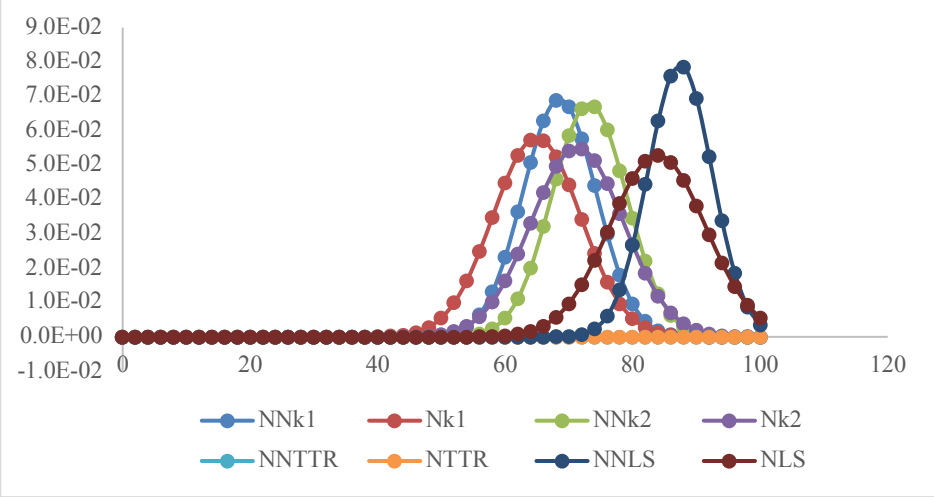


FIGURE 1. Distribution of the lexical indices

Note: *NNK1=First thousands of words in non-Anglophone texts; NNK2=Second thousands of words in non-Anglophone texts*
*NK1=First thousands of words in Anglophone texts; NNK2=Second thousands of words in non-Anglophone texts*
*NNTTR=Type-token radio in in non-Anglophone texts; NNLS=Lexical sophistication in in non-Anglophone texts*
*NTTR= Type-token radio in in Anglophone texts; NLS=Lexical sophistication in in Anglophone texts*

As shown in Table 2 above, the Anglophone texts informed having more lexical density than the other non-Anglophone ones; that is to say, RAA published in English-speaking contexts used somewhat more content words compared to their non-English-speaking counterparts. From the statistical analysis, the LD is considerably higher since it is over the 40 percent-scale for the LD in the written text compared to the function words (see Table 4). The figures related to lexical sophistication illustrate the range of academic vocabulary belonging to each discipline (see Table 5). The education abstracts of non-Anglophone journals informed a wide range of advanced words or specialized terminology (90.1), which, in some cases, are unique in each area and discipline. The gaining in this dimension also indicates that the radio of new words (type) to the total number of words (token) —lexical variation is unstable for short texts and can be affected by the length of texts. In other words, the longest the text, the lowest figure of LV, as in education texts (0.25). It is important to note that high lexical variation does not necessarily lead to a rich vocabulary if each

word formation is counted as a different word. For instance, when lemmatizing the education sub-corpora in both Anglophone and non-Anglophone groups, it was found that a headword (lemma) can have some constructions, as follows in the example in Table 3.

TABLE 3. Word formation in education RA abstracts

| Non-Anglophone groups | | | Anglophone groups | | |
|---|---|---|---|---|---|
| Lemma | Word | POS | Lemma | Word | POS |
| Write | prewriting | Noun | Real | realism | Noun |
| | rewrite | Verb | | realisms | " |
| | rewrites | " | | realist | " |
| | rewriting | " | | realistic | Adjective |
| | rewritten | Adjective | | realistically | Adverb |
| | unwritten | " | | realists | Noun |
| | writer | Noun | | realities | " |
| | writers | " | | reality | " |
| | writes | Verb | | unreal | Adjective |
| | writing | Noun | | unrealistic | " |
| | written | Adjective | | unrealistically | Adverb |
| | wrote | Verb | | unreality | Noun |
| Validate | invalidates | Verb | academy | academic | Adjective |
| | invalidating | " | | academically | Adverb |
| | validated | " | | academicals | Noun |
| | validates | " | | academician | " |
| | validating | " | | academicians | " |
| | validation | Noun | | academics | " |
| | validations | " | | academies | " |
| | | | | unacademic | Adjective |

Note: *POS=Part of the speech*

Table 4 and Figure 2 present the proportion of content and function words used in the two disciplines in Anglophone and non-Anglophone journals. As can be seen from the table and figure below, there is variation in the proportion of using content and functions words; thus, nouns are the highest, and pronouns are the lowest lexical features used in ANAJ. Comparing the proportion of lexical items with the total number of words, electronic texts of Anglophone and non-Anglophone groups had more content words (895 and 1013 tokens, respectively) than their counterparts. The table below illustrates the breakdown of the function words in the two disciplines. Electronics texts of non-Anglophone groups got considerably more coordinating conjunctions (59 tokens) than the other three data sets of RA abstracts.

TABLE 4. Lexical density in Anglophone and non-Anglophone written texts

| | | CONTENT WORDS | | | | FUNCTION WORDS | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | N | V | Adj | Adv | Pp | Dt | Cc | Pn |
| AJ-Education | *Token* | 479 | 150 | 107 | 53 | 173 | 94 | 47 | 34 |
| | *%* | (34) | (14) | (7) | (3) | (14) | (6) | (3) | (2) |
| NAJ-Education | *Token* | 521 | 159 | 129 | 22 | 223 | 188 | 42 | 29 |
| | *%* | (33) | (15) | (7) | (2) | (15) | (9) | (2) | (1) |
| AJ-Electronics | *Token* | 541 | 180 | 132 | 42 | 229 | 174 | 45 | 20 |
| | *%* | (32) | (13) | (7) | (1) | (15) | (11) | (3) | (1) |
| NAJ-Electronics | *Token* | 596 | 215 | 165 | 37 | 262 | 276 | 59 | 16 |
| | *%* | (30) | (15) | (8) | (1) | (15) | (12) | (2) | 0 |

Note: *N=nouns, V=verbs, Adj=adjective, Adv=adverb, Pp=preposition, Dt=determiner, Cc=coordinating conjunction, Pn=pronoun; NAJ= non-Anglophone journals; AJ=Anglophone journals*
*AJ= Anglophone journals; NAJ=non-Anglophone journals*

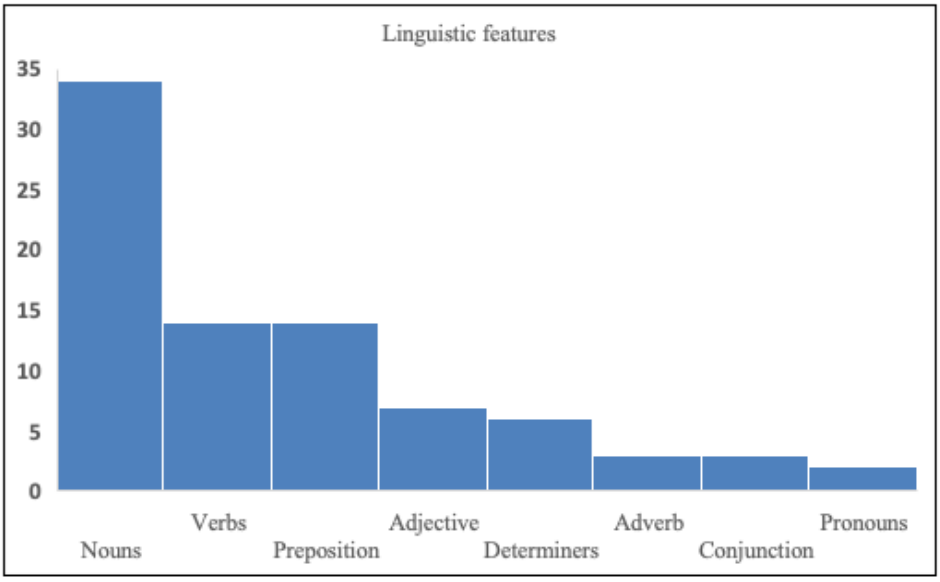


FIGURE 2. Content and function words in the whole corpora

As illustrated in Figure 2 above, the results bring considerable light to the content and function words used throughout the whole English written corpora. However, their frequency of occurrence decreases according to their usage in each discipline. That is, their incidence perhaps varies according to writers' language proficiency, writing skills, cultural background, and expertise in such disciplinary fields.

TABLE 5. Lexical richness in Anglophone and non-Anglophone written texts

| | Anglophone | | non-Anglophone | | *Correlation* | | |
|---|---|---|---|---|---|---|---|
| | Educ | Elect | Educ | Elect | LD-LS | LS-LV | LV-LD |
| Lexical Density | 0.65 | 0.63 | 0.57 | 0.58 | -.459 (.541) | | |
| Lexical Sophistication | 88.33 | 81.17 | 90.1 | 87.6 | | -.087 (.913) | |
| Lexical Variation | 0.29 | 0.27 | 0.25 | 0.29 | | | .473 (.527) |
| Words in text | 4061 | 5650 | 5176 | 4909 | | | |
| # of texts | 30 | 30 | 30 | 30 | | | |

Note: *Edu=education; Elect=electronics; LD=lexical density; LV=lexical variation; LS=lexical sophistication*

Table 5 describes the lexical statistic differences between lexical dimensions across disciplines and the context of publication. Abstracts published in ANAJ, like those of Tovar (2019), tended to use more words of lower frequency, more content words, and lower repetition of words to convey information content. Multiple running pair sample t-test made evident statistical differences between lexical dimensions, namely LS-LV (t (3)= -44.383, p= .000), LD-LV (t (3)= 19.539, p= .000), and LD- LS (t (3)= -44.029, p= .000), whose *Mean value* (-86.525; .332; 3.915) and *SD* (3.899; .0340; 3.915), respectively, raise diversities in lexical patterns. In general, the RA abstracts were considered relatively dense as they contained many lexical words concerning the total number of words. *Lexical items* are the words that primarily convey information in a text.

Accordingly, following Laufer and Nation (1995), the fewer incidence of function words reflects the occurrence of more subordinate clauses and participial phrases, which are not lexis but structural characteristics of the writing composition.

TABLE 6. Three-word formation in Anglophone and non-Anglophone written text

| 1238 3-wd in native | Share 6 3-wd | 1409 3-wd in nonnative |
|---|---|---|
| 1. as well as | 1. the importance of | 1. percentage of coverage |
| 2. a result of | 2. data were collected | 2. the development of |
| 3. more likely to | 3. in this study | 3. a number of |
| 4. their perceptions of | 4. purpose of this | 4. in order to |
| 5. within their schools | 5. the purpose of | 5. of oral homework |
| 6. data as well | 6. were collected from | 6. of serious games |
| 7. data from number | | 7. to determine the |
| 8. emphasize bilingual education | | 8. the method of |
| 9. is possible to | | 9. applied for the |
| 10. in the context | | 10. based on a |

Data in Table 6 exemplify the first ten 3-word string types used differently, and the first six 3-word string types shared in Anglophone and non-Anglophone written sub-corpora. In the current analysis, three-word formation accounts for high-frequency levels for a variety of specialized discourse types, contextual situations, and forthcoming purposes rather than statistical significance. This is especially true considering that 3-word string types were recurrent in the abstracts along with the English sub-corpora when constructing the information content and communicative goals. The multiword formation cannot be taken as the main characteristic used by a particular speaker or writer, rather a pattern widely used within discursive communities.

## DISCUSSION

The contrastive-descriptive analysis of journal research article abstracts has become perhaps, to date, the database for various linguistic research to document compositional patterns in writing. It was observed in Table 1 that the abstracts published in Anglophone context (U.S.A) produce seven sentences (with a length of 21.28 words) per abstract whereas those of non-Anglophone contexts (Ecuador) include six sentences (with a length of 34.44 words). The mean length of abstracts in Anglophone and non-Anglophone journals is 178.07 and 168.27 words, respectively. Regarding the surface structure, the descriptive statistics showed that non-Anglophone written texts exceed the standard average of thirty 30 lengths of words in a sentence (Biber & Conrad, 2009). Since sentence length is a good indicator of writing quality, surpassing the average word count could make the text grammatically complex and somehow interfere with reading comprehension. Although such differences are far from having a general conclusion, we can infer that these variations may respond to journal guideline requirements and authors' linguistic and cultural backgrounds. Literature review on this area claims that vocabulary and language knowledge positively affect the proportion of appropriate lexico-grammatical choices, which in some way contributes to the comprehension and quality of writing (Ha, 2019; Hu & Nation, 2000; Douglas, 2012; Read, 2000).

Regarding journal publishers' requirements for manuscript submissions, the existence of different guidelines for writing RAAs, in which experienced and inexperienced authors introduce the RAs, influences the factual summary of the research study. It is because abstracts as a

"representation" (Bazerman, 1984, p. 58), "distillation" (Swales, 1990, p. 179), "crystallization" (Salager-Mayer, 1990, p. 367), or "summary" (Kaplan et al., 1994, p. 405) of the text, "give the reader an exact and concise knowledge of the full article" (Bhatia, 1993, p. 78). In this way, authors should present accurate content and structure, reducing complex terms into easy ones by cutting off redundant and superfluous information. In essence, it does more than simply provide the "gist of the article in a precise and maximally efficient way"; instead, it does include the effective synergy of lexical and grammatical choices (Ventola, 1994, p. 333).

Journal RA abstracts acknowledge having a more lexical variation, containing many academic words with a considerable proportion of first and second 1000 words and incidence of low-frequent words, including lexical items. These features and the high information content made abstracts of RA published in English-speaking journals tend to use a slightly higher proportion of concrete words (M=0.637 with 56% frequency) than in non-English-speaking journals (M=0.597 with 55% frequency). The fact that the indices of the two English sub-corpora are slightly different announced that these abstracts are lexically dense. Additionally, most RAAs advertise a high level of lexical sophistication across the two disciplines between Anglophone and non-Anglophone groups. Nonetheless, the high frequency of lexical sophistication in abstracts of non-Anglophone journals (M=87,420) highlights that these texts incorporate more advanced vocabulary or specialized terminology than those of Anglophone journals (M=83.902).

As the lexical variation is based on the length of the text, the mean values of LV seem to be similar between abstracts in Anglophone (0,624) and non-Anglophone journals (0.601, respectively). However, further analysis let out far-reaching variation across disciplines and publication contexts; thus, education (TTR=0.29) and electronics (TTR=0.27) published in English-speaking contexts differ from those of (TTR=0.25; TTR=0.29) published in non-English-speaking contexts. It is because the measure is sensitive to the word count where word families are differentiated, identifying their code meaning and base form. Each word-formation then is not counted as a different word. Consequently, a text that showed many derived word forms of a few families could not be treated like a text that used many different word families (Laufer & Nation, 1995). This is because LV distinguishes how well a writer can produce texts with the vocabulary s/he knows; rather than what kind of words he knows. Following Connor (2004), such lexical variation may respond to authors' different discourse community practices and lingua-cultural conventions, which shape the context of the content. That is, the frequency of using standard and academic words may differ throughout the corpora since not all the most frequently occurring words listed as academic could appear in writing compositions (Waluyo & Kakoko, 2021).

The word family analysis, as a unit of comparison, between Anglophone and non-Anglophone written texts, education RAA inform that 126 families (786 repeated families) are shared in both texts, which resulted in 35.29% and 61.79% of coverage and text comprehensibility. Hence, 289 families (520 tokens) and 286 families (486 tokens) are the figures for unique frequencies in texts published in English-speaking and non-English-speaking contexts. What stands out from this analysis is that Anglophone texts produced 425 unshared families with an academic word-level of 82 families, 104 types, and 162 tokens. It also includes 102 (tokens) off-list words as proper nouns, capped at the mid-sentences (e.g., the words Latino *–8 tokens* and California *–3 tokens*), as well as 44 compounds with 88 tokens (e.g., bridge programs, policymakers, and engaging teachers), as described in the excerpts *a* and *b*:

a) Few current research examines language instruction educational programs in states with the more recent growth of the *Latino* English learner population (...), we first examine variations in EAP participation across *California*'s public high schools and what accounts for the variations in the early years of the program.
b) This article examines the effect of summer *bridge programs*, in which students enrol in coursework prior to beginning their first full academic year. Results have important consequences for *policymakers* and education leaders designing and implementing Common Core State Standards (…), we found that highly engaging and less *engaging teachers* differed of reflectiveness.

The possible reason for such differences among these English academic texts is due to writers' different writing experiences. This finding leads to Laufer and Nation's (2001) idea that vocabulary richness in scientific discourse, be it in oral or written form, varies according to disciplinary fields, linguistic competence, and context of the publication, which in conjunction with the lexical dimensions, predict academic writing success. Accordingly, the organization of ideas and the flow of writing, in a certain way, are determined by writers' unawareness of the rhetoric, limited vocabulary, and lack of native-like fluency (Bahtiar et al., 2020), which leads to having a messy and unclear linguistic and academic production of texts. That is why differences emerge, particularly in lexical and grammatical patterning (Lin & Lin, 2019) across texts in both academic discourses.

## CONCLUSION

Vocabulary knowledge is a good component of measuring lexical richness since it is the basis for constructing sentences, paragraphs, and complete texts. Lexical richness stands for the relative proportion of words that experts and novice writers use from different frequency levels in written texts from various disciplines. Therefore, it is seen as a paramount constituent of writing quality in non-native (and even native) academic written texts. Research findings advise variation in the lexical richness across disciplines and publication contexts in the three different dimensions analysed. The higher figures on lexical variation (TTR) evidence the use of varied and large vocabulary; it is a good indicator that academic authors have been exposed themselves to a wide range of reading materials from different types of texts. Concerning the fact that abstracts of English-speaking journals use a fewer incidence of low-frequency words than those of non-English-speaking journals, slight differences between the figures of lexical density and sophistication emerge. First, abstracts that unfold fewer sophisticated or advanced words do not necessarily reflect a lack of lexical density (Šišková, 2012). This idea seems to correlate, in the current study, with the journal RA abstracts published in English-speaking contexts. The abstracts in AJ apparently produce fewer indices of academic words, but in terms of lexical density, they produce higher indices than the texts in non-English-speaking contexts.

Abstracts of research articles in both disciplines are relatively comparable because they represent the sample population of education and electronics texts in the proportion of vocabulary used throughout these texts. Research outcomes could help language instructors and outsiders to understand how abstracts of scientific papers published in different contexts address the linguistic features of the language in actual use, which to some extent, are governed by the context of the publication. In the present study, lexical variation and sophistication were found to be the most

influential factor contributing to describing differences in writing across the two disciplinary fields. Such lexical variation is somewhat correlated to Hyland (2004), who pointed out that scientific discourse follows the conventions and writing practices of those institutionalized discursive communities. Although the research reported variation in the lexical richness across academic texts, particularly in diversity and sophistication, generalities are limited by the sample size and research design. Thus, in further research, text quality should be evaluated by measuring the dimensions of lexical richness and taking into account an in-depth analysis of the writers' academic writing competence.

To conclude, the unusual or specialized words across the two disciplines are recognised as an indicator of how writers employ vocabulary in scientific research papers (Laufer and Nation, 1995). However, it is imperative to mention that "there is no one measure of lexical richness which would give perfect results; researchers tend to use several different measures to obtain more information" (Šišková, 2012, p. 31). In this way, the generalizability of the results is subjected to various factors behind vocabulary size and vocabulary use that could affect lexical richness in writing. This could include topic familiarity, writing skills, communicative purpose, linguistic competence, among others. For instance, as stated above, the mastery of the target language may directly or indirectly determine the amount of varied and sophisticated vocabulary used in scholarly writing compositions. The findings tend to alert novice writers, especially those from non-English backgrounds (Loan Nguyen, 2018), to be aware not only of linguistic patterning but also of the content and structure of academic texts.

Although the current study has some limitations, such as two disciplines and context of publication rather than writers' nativeness, the results are valuable for writers and researchers, particularly for inexperienced and novice writers; this sheds light on using automatic online tools, such as vocab-profile to gauge the vocabulary size and vocabulary used in their written compositions. As a result of this analysis, they will be aware of the tendency to repeat words; this somehow allows them to produce texts with a varied and extensive vocabulary. To put it in another way, speakers' broader vocabulary repertoire allows them to avoid repeating the use of vocabulary as a means of linking their ideas to a written text (Hung et al. 2021). Additionally, the findings suggest the necessity for corpus-based vocabulary guidelines, that is, carefully selecting the most familiar and recurring advanced words from authentic scholarly materials, non-native speakers of English not only improve writing quality but upturn reading comprehension. It is because knowing at least 4000 to 5000 words is the basis for understanding academic texts and 6000 to 8000 words for reading them.

Finally, the research findings attempt to contribute to genre analysis, academic writing and L2 writing instruction (Al-mudhaffari et al., 2019) so that writers of English as a Second or Foreign language can include appropriate lexical and grammatical entries when producing scholarly and non-scholarly texts. From this point of view, Abdullah et al. (2021) suggest that the teaching of the lexicon must be accompanied by syntactic structures; since the knowledge of the lexical elements implies knowing the semantic and syntactic structures of the language (Mohamed Sultan & Taha, 2018), which, to some extent, could impact the oral or written production of texts.

ACKNOWLEDGEMENTS

The article has not received any private or public funding. I would like to extend my grateful acknowledgement to the constructive comments sent by the anonymous reviewers in making this research stronger enough. I also thank Assoc. Prof. Dr. Fazal Mohamed Mohamed Sultan for reviewing and considering this manuscript for publication and for his timely and helpful e-mails that help follow journal policies and improve the quality of the current article. Last but not least, to some of my former students who participated in this research during the data collection and selection process.

REFERENCES

Aarts, B. (2014). *The Oxford dictionary of English grammar* (2nd ed.). Oxford University Press.

Abdullah, S., Abdul Aziz, R., & Kamaruddin, R. (2021). Lexical verbs in verb-noun collocations: Empirical evidence from a Malay ESL Learner Corpus. *3L: Language, Linguistics, Literature: The Southeast Asian Journal of English Language Studies, 27*(4), 144-156. http://doi.org/10.17576/3L-2021-2704-11

Al-mudhaffari, M., Hussin, S., & Ho Abdullah, I. (2019). Interaction in Academic L2 writing: An analysis of Interactional Metadiscourse Strategies in Applied Linguistics Research Articles. *3L: The Southeast Asian Journal of English Language Studies*, *25*(3), 16-32. http://doi.org/10.17576/3L-2019-2503-02

Arnold, I. V. (1986). *The English word, Department of English Philology*, Orenburg State Pedagogical Institute named after V.P. Chkalova. https://bit.ly/3znP1Fc

Bahtiar, S., Samad, A., Razali, A., & Wei, H. (2020). A corpus analysis of frequently occurring words and their collocations in high-impact research articles in education. *3L: The Southeast Asian Journal of English Language Studies,* 26(4), 99-113. http://doi.org/10.17576/3L-2020-2604-08

Bazerman, C. (1984). Modern evolution of the experimental report in Physics: Spectroscopic articles in Physical Review, 1893-1980. *Social Studies of Science*, *14*(2), 163-196. http://www.jstor.org/stable/284651

Bhatia, V. K. (1993). *Analyzing genre: Language use in professional settings*. Longman Publishing.

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge University Press.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Cobb, T. (2006). Complete lexical tutor. http://www.lextutor.ca

Colman, A.M. (2015). *Dictionary of Psychology* (4th ed.). Oxford University Press

Connor, U. (2004). Intercultural rhetoric research: Beyond texts. *Journal of English for academic purposes*, *3*, 291-304. https://doi.org/10.1016/j.jeap.2004.07.003

Constitución del Ecuador, ciencia, tecnología, innovación y saberes ancestrales. Asamblea Constituyente. [Constitution of Ecuador, Science, technology, innovation and ancestral knowledge.] (2008). Constituent Assembly. https://bit.ly/3iANdCY https://bit.ly/3Lc16DN

Djiwandono, P. I. (2016, 21 January). Lexical richness in academic papers: A comparison between students' and lecturers' essays. *Indonesian Journal of Applied Linguistics,* *5*(2), 209-216. http://www.doi.org/10.17509/ijal.v5i2.1345

Douglas, S. R. (2012). Non-Native English-Speaking Students at University: Lexical Richness and Academic Success. (Unpublished dissertation). Universitas Calgary, Alberta. http://hdl.handle.net/1880/48195

Douglas, S. R. (2019). The productive lexical profile of an undergraduate non-credit academic writing course. In A. Tsedryk & C. Doe (Eds.*)*, *The description, measurement, and pedagogy of words* (pp. 142–172). Cambridge Scholars Publishing. https://bit.ly/3NWo18v

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics*, 35(3), 328-356. https://doi.org/10.1093/applin/amt016

Gardner, D. (2013). Exploring vocabulary: Language in action. Routledge. https://bit.ly/2UBX4vX

Ha, H. S. (2019). Lexical richness in EFL undergraduate students' academic writing. *English Teaching*, *74*(3), 3-28.

Halliday, M. A., & Hassan, R. (1976). *Coherence in English*. Longman.

Harvey, B. (2000). *Stat Trek Random Number Generator*. Georgia Institute of Technology, Retrieved February 20, 2021, from http://stattrek.com/Tables/Random.aspx

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, *13*(1), 403-430.

Hung, B., Quang, H., Nguyen, L., & Thi Viet, N. (2021). A Cross-Linguistic Approach to Analysing Cohesive Devices in Expository Writing by Asian EFL teachers. *3L: Language, Linguistics, Literature: The Southeast Asian Journal of English Language Studies*, *27*(2), 16-30. http://doi.org/10.17576/3L-2021-2702-02

Hyland, K. (2004). *Disciplinary discourses. social interactions in academic writing* (2nd ed.). Michigan Classics Edition.

Instituto Nacional de Estadística y Censos: Encuesta nacional de nmpleo, Desempleo y subempleo (ENEMDU). [National Institute of Statistics and Censuses: National survey of employment, Unemployment and underemployment (ENEMDU)], (2018). https://www.ecuadorencifras.gob.ec/enemdu-2018/

Kaplan, R., Cantor, S., Hagstrom, C., Kamhi-Stein, L., Shiotani, Y., & Zimmerman, C. (1994) On abstract writing. *Text & Talk - Interdisciplinary Journal for the Study of Discourse*, *14*(3), 401-426. https://www.degruyter.com/journal/key/text/14/3/html

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? In S. Foster-Cohen & A. Nizegorodcew (Eds.), *EUROSLA yearbook* (pp. 7-28). John Benjamins.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing.* 21(2), 202-226. https://doi.org/10.1191/0265532204lt277oa

Lee, J. (2018). A Coh-metrix analysis of lexical, syntactic and discourse aspects in the newspaper articles of Korean and British university students. *Modern English Education*, *19*(4), 17-26. https://doi.org/10.18095/meeso.2018.19.4.17

Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward*. Language Teaching Publications.

Lin, C., & Lin, Y. (2019). Grammatical and lexical patterning of make in asian learner writing: A corpus-based study of ICNALE. *3L: The Southeast Asian Journal of English Language Studies, 25*(3), 1-15. http://doi.org/10.17576/3L-2019-2503-01

Lipka, L. (2002). *English lexicology: Lexical structure, word semantics & word formation*. Tubingen, Germany. Gunter Narr Verlag Tubingen. https://bit.ly/3AkmXnr

Loan Nguyen, T. (2018). Rhetorical structures and linguistic features of English abstracts in Thai Rajabhat University journals. *3L: The Southeast Asian Journal of English Language Studies, 24*(4), 71–84. http://doi.org/10.17576/3L-2018-2404-06

Mohamed Sultan, F.M. & Taha, M. (2018). The syntax of applicative Constructions in spoken Sudanese Arabic. *3L: Language, Linguistics, Literature*, *24*(2), 143-153.

McArthur, T., Lam-McArthur, J. & Fontaine, L. (2018). *The Oxford companion to the English language* (2nd ed.). Oxford University Press.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge University Press.

Pho, P.D. (2008). "Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance". *Discourse Studies*. *10*(2), 231-250, https://doi.org/10.1177/1461445607087010

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Romero, N. (2020). Lexical richness and readability in English degree students' writing. [Unpublished master's thesis]. Graduate Department. Technical University of Cotopaxi. Latacunga-Ecuador.

Salager-Mayer, F. (1990). Discoursal flaws in medical English abstracts: a genre analysis per research and text type. *Text.* 10(4), 365-384. https://doi.org/10.1515/text.1.1990.10.4.365

Sánchez, P. (2020). Lexical bundles and moves in research article abstracts of different disciplines. *ODISEA: Revista de estudios ingleses*. *21*, 71-87. https://doi.org/10.25115/odisea.v0i21.4130

Sayfouri, N. (2010). *SFL and ESP genre analysis of English research articles in Iranian and English medical journals: A contrastive study* [Published Doctoral Thesis]. University of Tehran, Faculty of Foreign Languages and Literatures. https://doi.org/10.13140/RG.2.1.4084.7446

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *University of Reading Language Studies Working Papers*, *4*, 26–36. https://bit.ly/3A1E3FT

Strang, B. (1968). *Modern English structure, 2nd ed.* Edward Arnold (Publishers) Ltd.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tankó, G. (2017). Literary research article abstracts: An analysis of rhetorical moves and their linguistic realizations. *Journal of English for Academic Purposes*. *27*, 42-55. https://doi.org/10.1016/j.jeap.2017.04.003

Tovar, R. (2017). Vocabulary knowledge in the production of written texts: a case study on EFL language learners. *Revista Tecnologica ESPOL*. 30(3), 89-105. https://bit.ly/3Ju2v8g https://bit.ly/3N9kPWB

Tovar, R. (2019). Rhetorical move structure in abstracts of research articles published in Ecuadorian and American english-speaking contexts. *Arab World English Journal*, *10*(4), 74-87. https://dx.doi.org/10.24093/awej/vol10no4.6

Van Bonn, S., & Swales, J.M. (2007). English and French journal abstracts in the language sciences: Three exploratory studies. *Journal of English for Academic Purposes*, *6*(2), 93-108, https://doi.org/10.1016/j.jeap.2007.04.001

Ventola, E. (1994). Abstracts as an object of linguistic study. In S. Cmejrkova, F. Danes & E. Havolova (Eds.), *Writing vs. speaking: Language, text, discourse, communication* (pp. 333-352). Gunter Narr.

Waluyo, B., & Kakoko, R. (2021). Academic vocabulary used by high school students in essays and its relation to English proficiency. *3L: Language, Linguistics, Literature. The Southeast Asian Journal of English Language Studies. 27*(4),157-172. http://doi.org/10.17576/3L-2021-2704-12